Condamines, A. & Rebeyrolle, J. (1997c). Utilisation d'outils dans la constitution de Bases de Connaissances Terminologiques : expérimentation, limites, définition d'une méthodologie. In Actes *1ères Journées Scientifiques et Techniques du Réseau Francophone de l'Ingénierie de la Langue de l'AUPELF-UREF*, Avignon, avril 1997, pp.529-535.

Utilisation d'Outils dans la Constitution de Bases de Connaissances Terminologiques : Expérimentation, Limites, Définition d'une Méthodologie

Anne Condamines (CNRS), **Josette Rebeyrolle** (Université Toulouse Le Mirail)
Equipe de Recherche en Syntaxe et Sémantique (ERSS), Maison de la Recherche
5 Allées Antonio machado, 31058 Toulouse cedex.
tel : 05 61 50 36 02 fax : 05 61 50 46 77
e-mail : condamin@irit.fr rebeyrol@cict.fr

Résumé

De nombreux laboratoires, publics ou privés, travaillent à la construction d'outils pour aider à l'acquisition de données terminologiques à partir de textes. On peut cependant regretter qu'ils soient encore très peu utilisés ce qui rend difficile l'évaluation de leur pertinence réelle. Du fait de notre engagement dans la constitution de bases de connaissances terminologiques sur des corpus d'entreprises, nous devons définir des méthodes qui intègrent le mieux possible l'utilisation d'outils d'analyse de corpus. Notre travail s'organise en deux parties, d'une part, recensement et test d'outils, d'autre part, intégration des résultats fournis par ces outils à une méthode de constitution de terminologies. Nous proposons une méthode précise d'utilisation de plusieurs outils : Lexter, Hyperbase, SATO afin de constituer une base de connaissances terminologiques. L'article décrira l'expérimentation en cours ; il s'attardera surtout sur la présentation de la méthodologie que nous mettons en oeuvre qui consiste en une intégration de résultats fournis par différents outils guidée, à la fois par l'objectif, constituer une Base de Connaissances Terminologiques (BCT), et par les connaissances que nous avons sur le fonctionnement de la langue.

1. Introduction

L'utilisation effective d'outils pour la constitution de Bases de Connaissances Terminologiques (BCT) ne va pas sans poser de problèmes, d'autant plus difficiles à identifier que les outils dédiés à la terminologie sont souvent des prototypes qui n'ont pas été évalués sur des projets effectifs de constitution de terminologies.

Depuis plusieurs années, nous travaillons à la définition de méthodes de constitution de BCT[1] ce qui nous amène à tester des outils et à essayer d'intégrer les résultats qu'ils produisent dans notre démarche d'analyse de textes. Or, l'utilisation de ces outils entraîne des problèmes qui, selon nous, linguistes utilisateurs de ces outils, peuvent avoir pour origine, d'une part, le fait que la réflexion théorique sous-jacente à la constitution n'est souvent pas suffisamment avancée, et d'autre part, le fait que les outils, définis pour un résultat ponctuel, laissent peu de marge de manoeuvre à l'utilisateur.

Notre présentation s'organisera en trois étapes. Nous indiquerons d'abord quel objectif nous poursuivons lors de l'analyse de textes (présentation du modèle de BCT et des tâches qu'il amène à effectuer), ensuite nous exposerons notre méthode de travail qui allie analyse linguistique et utilisation de différents types d'outils, puis nous proposerons, à travers un exemple, une illustration de notre démarche.

2. Constitution de Bases de Connaissances Terminologiques

La constitution d'une base de connaissances terminologiques suppose qu'un ensemble d'éléments soient clairement établi : un corpus doit servir de référence et de matériau de travail, un modèle de données doit avoir été défini, les étapes d'analyse du corpus doivent avoir été identifiées, le type de résultats fournis par les outils doivent avoir été analysés et intégrés, les connaissances linguistiques qui sont mises en oeuvre ou crées doivent avoir été déterminées. Dans ce processus, on le voit, la mise en oeuvre des outils constitue un des éléments qui est contraint par d'autres : la nature des données à rechercher et les connaissances linguistiques utilisées.

2.1 Le modèle de BCT

Pour rendre compte du fonctionnement de la terminologie d'un domaine, nous avons retenu un certain nombre de données nécessaires (terme (T), concept (C), lien terme/concept (LT/C) et texte)) qui sont organisées dans le modèle suivant (Condamines & al., 1993) (Séguéla & al., 1997) :

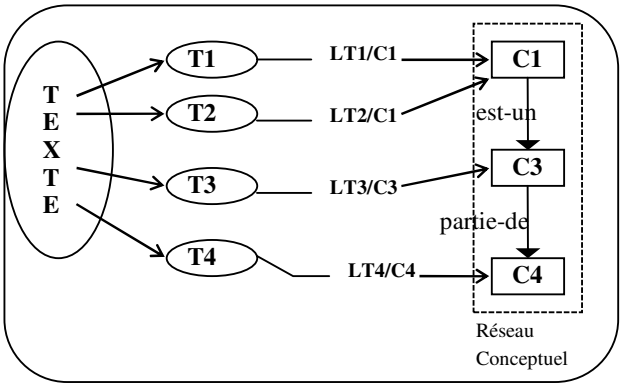

Figure 1 : Les différents composants de la BCT

Les quatre champs du modèle contiennent respectivement les informations suivantes :

[1] Anne Condamines et l'ERSS ont reçu, en 1994, le prix ANVIE/CNRS de valorisation de la recherche en sciences humaines dans les entreprises pour la constitution de BCT.

- le « terme » comporte les données proprement linguistiques (nature et genre, variantes de formes : sigles, abréviations, etc.) ;
- le « concept » contient les données qui concernent le concept dénommé par le terme, sous la forme d'une définition et de relations sémantiques explicites. Le choix des relations n'est pas contraint et seule la relation taxinomique est réellement formalisée ;
- le « lien terme/concept » renferme des informations sur les contraintes de la validité d'un terme pour dénommer tel concept ; il s'agit, par exemple, de standards ou de normes ou bien de certains locuteurs (tel ou tel département à l'intérieur d'une entreprise, tel ou tel métier, etc.) ;
- le « texte » permet de rendre compte des liens entre un terme et ses occurrences dans un corpus représentatif d'un domaine ou d'une tâche que la BCT cherche à représenter. Dans le cas d'un terme dénommant plusieurs concepts (polysémie ou homonymie), les occurrences sont sélectionnées en fonction de chaque concept.

2.2 Les tâches à effectuer

Pour compléter ce modèle, on distingue quatre types de tâches à réaliser (Condamines, 1996a) :
- repérage des termes : par exemple, à partir d'une liste de candidats termes ;
- repérage d'équivalents de forme ou de variantes de termes : il s'agit des sigles ou des abréviations, voire de variantes plus sophistiquées, comme *IVA moyenne* pour *segment moyen de l'IVA*, ce sont des groupes de mots qui apparaissent dans des contextes sémantiquement proches et qui entretiennent une parenté morphologique ;
- repérage des équivalents de contenu pour des formes différentes (« synonymes »), c'est-à-dire des relations entre termes, mots ou syntagmes, qui apparaissent dans des contextes sémantiquement proches sans avoir aucune parenté morphologique (*capteur, senseur, détecteur*, dans la terminologie de Matra Marconi Space) ;
- repérage des relations entre concepts : pour des couples de termes, repérage de contextes différents mais sémantiquement proches.

Le travail d'analyse de textes, sur lequel se fonde le recueil des données nécessaires à une BCT, fait appel à des connaissances sur le fonctionnement de la langue ; autant que faire se peut, cette mise en oeuvre doit être contrôlée par le linguiste. En effet, si l'on vise à définir une méthode qui soit réutilisable voire enseignable, on ne peut se contenter d'une approche intuitive. Nous montrerons dans la partie suivante comment ces connaissances sont activées à chaque étape du recueil. On montrera que suivant les cas, la connaissance sur la langue de référence permet soit de mettre en évidence un fonctionnement régulier mais implicite dans les corpus, soit de mettre en évidence un fonctionnement déviant par rapport au fonctionnement attendu. Dans tous les cas, l'analyse de textes spécialisés est faite par comparaison avec un système considéré comme standard et stabilisé.

3. Méthode de constitution de BCT

Une fois établi le type de données qui vont être recherchées dans les corpus, on peut mettre en place une méthode de recueil de ces données et rechercher les outils qui assisteront le mieux cette démarche.

3.1 Les outils

Deux types d'outils peuvent être envisagés, des outils construits pour l'extraction de données terminologiques et des outils d'analyse de textes au sens large.
Le premier type d'outil propose soit des candidats termes, soit des candidates relations conceptuelles. Que la démarche soit de type ascendant (les données sont remontées du corpus comme dans les outils de type statistique) ou de type descendant (les données sont définies a priori et recherchées dans le corpus comme dans (Hearst, 1992)), tous ces outils engendrent du silence ou du bruit. Le linguiste-terminologue doit alors intervenir pour, en fonction des données qu'il recherche, supprimer les données inadéquates et récupérer les données oubliées. A ce moment-là, il a besoin d'outils très souples qui lui permettent des interactions rapides en fonction des résultats proposés. Alors, l'utilisation du second type d'outils est incontournable. Le second type d'outils n'est pas dédié à une recherche précise mais propose un ensemble de fonctionnalités que l'utilisateur doit mettre en oeuvre en fonction de ses besoins. Les concordanciers sont très caractéristiques de ce type d'outils.

Dans l'expérimentation en cours, nous utilisons un outil spécifiquement conçu pour la terminologie, Lexter ; nous avons retenu également deux autres types d'outils : Sato et Hyperbase.
- Lexter : le logiciel LEXTER (Logiciel d'EXtraction de TERminologie) a été conçu par Didier Bourigault (Bourigault, 1995) au sein de la Direction des Etudes et Recherches d'EDF pour aider à l'extraction de données terminologiques. Lexter utilise en entrée des corpus de textes techniques d'un domaine quelconque qu'il traite au moyen d'une analyse syntaxique automatique partielle. Sur la base de patrons morphosyntaxiques qui permettent de délimiter les frontières de groupes nominaux, le logiciel fournit en sortie une liste d'unités terminologiques candidates susceptibles de représenter les concepts du domaine étudié.
- Sato : le logiciel SATO (Système d'Analyse de Textes par Ordinateur) a été conçu par Jean-Guy Meunier et développé par François Daoust (Daoust, 1992). L'accès au texte s'effectue au moyen de concordances, c'est-à-dire de recherche de l'ensemble des occurrences d'un mot dans chacun de ces environnements contextuels. La grande originalité de Sato, par rapport aux autres concordanciers, réside dans le fait qu'il permet d'ajouter des propriétés aux mots ou aux segments textuels. Notons que l'ajout de propriétés peut être le résultat d'une opération automatique de projection d'une base de données lexicales[2] sur le lexique qui permet d'attribuer à chaque forme l'une des valeurs grammaticales déclarées (nom, verbe conjugué, pronom relatif, etc.). Par exemple, une commande comme
$(ment$,tion$,age$)
.2(de,d',du,des).2$*gramr=Ncom*.* permet de ramener la plupart des cas de déverbaux (ou nominalisations) qui sont suivis (dans un intervalle de deux mots après) de la

[2] Il s'agit de la Base de Données Lexicales (BDL) développée par L. Dupuy (Université de Québec à Montréal). L'algorithme de la BDL repose sur la comparaison des chaînes du lexique du corpus de texte à celles contenues dans les dictionnaires. Si les chaînes sont identiques, la chaîne du lexique reçoit la catégorie associée à la chaîne du dictionnaire.

préposition *de* puis (dans un intervalle de deux mots après) d'un nom commun. On trouverait ainsi : *intégration réussie des composants, prototypage d'un logiciel*.
- Hyperbase : construit en 1989 par Etienne Brunet (Brunet, 1995), Hyperbase « répond, par les méthodes de l'hypertexte, aux besoins classiques du traitement automatique des textes : concordances de type Key Word In Contexte (KWIC), index sélectifs ou systématiques, dictionnaires de fréquences, ..., cooccurrences, filtrage et masquage de mots, ..., recherche de parties de mots ». La fonctionnalité qui nous a particulièrement intéressées est celle qui permet de faire une comparaison, sous forme d'écart réduit, avec le corpus du Trésor de la Langue Française (TLF) de 70 millions de mots.

3.2 La méthode

Pour retrouver les informations associées à chacune des entités du modèle de BCT proposé, nous avons élaboré une méthode linguistique assistée par les outils que nous venons de citer. Nous appuyant sur les deux données principales du modèle : le terme et le concept, nous montrerons dans ce qui suit comment peuvent être repérées les informations nécessaires à leur description.

3.2.1 Recueil des données attachées au « terme »

a) Repérage de termes candidats

L'identification des termes est accélérée si l'on a recours à un extracteur de candidats termes comme Lexter. Néanmoins, les résultats que fournis Lexter ne sont pas utilisables directement car ils contiennent trop de bruit. Ainsi, ils demandent à être triés sur la base de critères linguistiques afin de constituer une liste de termes candidats, c'est-à-dire de chaînes de caractères dont on peut penser qu'elles sont des termes, avant toute analyse en corpus. Il s'agit, en effet, de spécifier des critères linguistiques relativement stables qui permettent d'exclure les unités qui constituent incontestablement du bruit et qui multiplient inutilement les données à analyser en contexte. Ainsi, nous proposons de rejeter les unités qui répondent à l'un des critères syntaxiques ou sémantiques énumérés dans ce qui suit.

On élimine, dans une première étape de filtrage, les candidats qui ne peuvent pas être des termes pour des raisons syntaxiques, qu'il s'agisse d'une erreur de découpage syntaxique effectué par Lexter (*présume pas de la méthodologie de développement utilisée*) ou d'une erreur de découpage due à une confusion entre forme verbale et forme nominale (la forme verbale utilisée est considérée par Lexter comme une forme nominale) : *offre*, *demande*. Seul un examen en contexte permet d'identifier cette erreur. C'est avec Sato que s'effectue cette confirmation. Les formes non-terminologiques sont également supprimées, il s'agit de structures qui sont syntaxiquement correctes mais qui ne sont pas terminologiques (*est-il, a-t-il, choix entre*).

Dans une seconde étape, on retire les candidats qui ne peuvent pas être des termes pour des raison sémantiques. Ainsi, certaines formes sont considérées comme étant trop générales, c'est le cas de certains groupes nominaux (*complément du paragraphe, Synthèse du projet*), de certaines locutions prépositionnelles (*à la suite de, à l'issue de*), de certaines locutions adverbiales (*coup par coup*), de certains candidats termes dont la tête joue un rôle de déterminant, et non pas de nom au sens plein (*famille, ensemble*), des formes contenant un anaphorique ou un cataphorique, dont l'interprétation est directement liée au contexte (*phase suivante de développement*), et également des formes contenant un déictique, dont l'interprétation est liée à la situation d'énonciation (*mise à jour du présent document*), et enfin des formes contenant un adjectif qualificatif trop vague ou trop général (*alphabétique, classique, concret, constitutif*).

b) Définition des limites terminologiques du domaine

Une des difficultés dans la constitution de BCT est liée à la clôture du domaine : aux marges du réseau conceptuel, comment définir ce qui est terme et ce qui n'est pas terme? Dans cet objectif, nous nous appuyons sur les résultats fournis par Hyperbase. En utilisant la fonctionnalité d'Hyperbase qui permet une comparaison avec le corpus du TLF, on peut, en effet, obtenir les mots qui, dans le corpus que l'on étudie, apparaissent en excédent ou en déficit. Nous nous intéressons particulièrement aux mots qui sont en excédent, en faisant l'hypothèse qu'ils vont sans doute jouer un rôle important dans le domaine et donc qu'on va pouvoir les envisager comme constituant une première délimitation du domaine. (Cette hypothèse est encore à l'étude, il faut donc la considérer avec précaution).

c) Repérage des équivalents de forme ou variantes de termes

Partant de la liste des candidats proposés par Lexter, on peut également avancer un certain nombre de critères de conservation de certains termes. Ainsi, nous avons défini un ensemble de critères morphosyntaxiques de conservation. Ces critères concernent les cas où existe une équivalence entre candidats termes, qui manifeste la présence d'un concept unique. On peut proposer des équivalences entre des candidats morphologiquement proches, comme : *outil de GL du projet* et *outil génie logiciel du projet*. Le logiciel Lexter propose d'ailleurs des hypothèses d'équivalence entre candidats termes, en repérant des variations de la détermination : T1 dét T2 vs T1 dét_Ø T2[3], comme dans l'exemple suivant : *état de configuration logiciel* vs *état de configuration du logiciel*. De même, on rapproche assez rapidement des équivalences entre des formes siglées et les formes développées correspondantes : *PDL = Plan de Développement Logiciel, DCP = Dossier de Conception Produit*.

Néanmoins, il s'agit d'autant d'hypothèses d'équivalence qu'il est nécessaire de valider avec les experts du domaine.

d) Repérage de termes qui n'ont pas de base nominale

Hyperbase est utilisé dans l'objectif de travailler sur les concepts qui renvoient à des actions ce qui, indirectement, permet d'accéder aux verbes qui ne sont que rarement proposés par les outils d'extraction de candidats-termes. Notre effort porte plus particulièrement sur les déverbaux, il s'agit des noms morphologiquement et sémantiquement associés à un verbe (*gestion/gérer*). Cette étude nous permet de faire une analyse approfondie des actions

[3]Symboles utilisés :
- dét : déterminant (dét_déf. : défini / dét_indéf. : indéfini / dét_Ø : absence de déterminant),
- T1 / T2 : termes ou " termes candidats " de Lexter.

spécifiques d'un corpus. La première liste obtenue grâce à Hyperbase fait actuellement l'objet d'une étude linguistique détaillée (à paraître) qui vise à définir des tests linguistiques permettant de sélectionner, de façon méthodique, les déverbaux qu'il convient de conserver comme termes.

Cette première série d'études permet d'acquérir une liste de termes candidats et de commencer à identifier des variantes de termes (sigles, ellipses) et des termes équivalents (validés par un expert).

3.2.2 Recueil des données attachées au « concept »

Le champ « concept » du modèle de BCT est identifié en particulier par les relations conceptuelles qui lient l'ensemble des concepts d'un domaine ; ce sont ces données qu'il faut retrouver dans le corpus étudié.

a) Repérage des équivalents de contenu pour des formes différentes ("synonymes")

Ces équivalences peuvent être tirées de l'examen des candidats termes de Lexter, si l'on applique certains critères sémantiques. Ces critères sémantiques permettent de poser l'existence de paradigmes d'équivalents parmi les expansions. Ainsi, des structures du type : T1 + Adj1, T1 + Adj2, dans lesquelles Adj1 et Adj2 sont synonymes en langue, peuvent être considérées comme équivalentes. Ce fonctionnement régulier est considéré comme un argument pour conserver les candidats termes : *petit projet* vs *projet de petite taille*. De la même façon, une opposition parmi les expansions est un argument pour retenir les candidats-terme : *conception générale* vs *conception détaillée*.

Par le biais de la notion d'équivalence, on accède au niveau conceptuel puisque deux termes sont équivalents s'ils renvoient au même concept.

b) Repérage des relations conceptuelles

Un examen des candidats termes de Lexter peut aussi permettre de mettre au jour certaines relations taxonomiques potentielles entre des candidats. En effet, les paradigmes proposés par Lexter sous la forme tête (T) et expansion (E) permettent de faire l'hypothèse qu'il existe une relation de hiérarchie entre T et T + E (ex : *test, test d'acceptation, test de qualification, test de validation, test de recette*, etc.).

Pour aller plus avant dans la détermination des relations conceptuelles, une analyse des corpus est indispensable. Dans la plupart des cas, c'est l'utilisation de fonctionnements réguliers qui va donner au linguiste une stabilité pour lui permettre d'expliciter des phénomènes implicites dans les textes spécialisés. Ainsi, lors de la construction de terminologies, le recours à des marqueurs syntactico-sémantiques permet de mettre en évidence des relations conceptuelles. Par exemple, une même relation de cause va s'exprimer différemment dans les trois phrases suivantes : *un infarctus du myocarde par sténose de l'IVA, la sténose de l'IVA est responsable de l'infarctus du myocarde, un IDM en relation avec une sténose de l'IVA*. Or, ces indices (que nous appellerons désormais marqueurs) de la cause sont utilisables quels que soient les domaines et permettent à un non-spécialiste de mettre au jour des portions entières de réseaux conceptuels, l'expert n'étant sollicité que pour confirmer les hypothèses. En revanche, il se peut que certaines marqueurs soient spécifiques à certains domaines ou même que certaines relations n'apparaissent que dans certains domaines (par exemple, la relation *est symptôme de* dans le domaine médical).

Pour repérer les relations et définir le réseau conceptuel, nous proposons de recourir à l'utilisation de Sato.

Notre étude de différents corpus et, plus particulièrement, notre analyse des marqueurs de relations définitoires (contextes très riches dans l'expression de relations conceptuelles (Rebeyrolle, 1996)) nous permet de caractériser le fonctionnement des marqueurs dans les corpus spécialisés.

On peut distinguer des marqueurs indépendants d'un corpus particulier (Borillo, 1996), ce qui ne signifie pas qu'ils sont systématiquement utilisés. Ainsi, dans le corpus en cours d'étude (cf. note 5), on retrouve certaines structures. Il s'agit de structures prédicatives, du type : dét_déf. **T1** est dét_indéf T2[4] (*Le Plan de Développement Logiciel est un document*) ; de structures appositives, soit strictement apposées, du type : dét_indéf. T2, dét_déf. **T1**, soit parenthétiques : dét_déf T2 (dét_∅ **T1**) (*Le produit logiciel correspond à une transformation (génération) d'un Etat de Configuration donné*) ; et également d'autres structures signalant un emploi autonymique, telle que : dét_déf **T1** c'est-à-dire dét_déf T2 (*tests à effectuer à l'issue de la phase de spécification, c'est-à-dire le plan de validation du logiciel*).

Par contre, dans ce même corpus, on ne retrouve pas les structures signalant l'introduction d'une définition, comme : on appelle dét_∅ **T1** dét_ indéf. T2, (structure pourtant très fréquente dans des corpus de type didactique) et dét_T2 ou dét_∅ **T1**, structure utilisée dans certains corpus pour introduire une équivalence entre des termes

On distingue, d'autre part, des marqueurs qui sont dépendants d'un corpus (cf expérimentation décrite en 4). Enfin, on peut trouver des marqueurs spécifiques pour une relation non identifiée.

Notre travail sur les corpus consiste à utiliser des connaissances que nous avons, a priori, sur les marqueurs de relations, soit pour les retrouver dans les corpus, soit pour les adapter au corpus, soit pour en découvrir de nouveaux. Dans cet objectif, une interaction permanente avec le logiciel nous permet de prendre en compte les résultats immédiatement et d'ajuster nos patrons de fouille, au fur et à mesure des réponses fournies. Ainsi, à partir de marqueurs considérés comme généraux (indépendants du domaine), nous définissons une première série de patrons de fouille que nous testons sur le corpus. A partir des résultats produits, nous adaptons les patrons pour les rendre plus pertinents, c'est-à-dire, pour limiter les bruits et les silences. Cette démarche est présentée dans le paragraphe suivant.

4. Expérimentation

Il s'agit de présenter pas à pas une utilisation de Sato pour le repérage d'informations conceptuelles, en montrant comment à partir d'une structure très générale, on peut

[4] L'utilisation du gras signale l'emploi autonymique du terme.
Il faut noter que l'on regroupe dans un même schéma général des structures qui peuvent subir des modifications syntaxiques, selon qu'elles se trouvent dans une relative, qu'elles sont précédées d'un présentatif, etc.

affiner une requête pour mettre en évidence des phénomènes propres à un corpus.
Nous traiterons plus précisèment de l'organisation temporelle des phases constitutives du processus de développement de logiciels scientifiques et techniques dans un corpus d'EDF[5] pour montrer comment, à partir d'un ensemble de marqueurs potentiels, on peut repérer d'autres marqueurs spécifiques à ce domaine.
Notons que nous ne détaillons pas ici la première phase, qui mène de la liste des candidats-termes proposés par Lexter à une liste de termes-candidats. Signalons simplement que l'application des critères présentés en 3.2.1 nous a permis de réduire la liste de 5875 candidats termes (pour un corpus d'environ 50 000 mots) à 1516 termes candidats, soit une réduction de 74 %. Nous nous intéressons ici à la deuxième étape du processus qui conduit à mettre en place un réseau relationnel.
Une étude de la relation *partie-de* nous a permis de mettre en évidence l'importance de la notion de *phase* pour le corpus. Ainsi, on a montré que l'organisation d'un processus en phases peut être modélisée grâce à une relation de type *partie-de*, accompagnée de relations comme *précède* ou *se superpose à* qui permettent d'ordonner ce type de parties (cet ordonnancement n'étant généralement pas nécessaire lorsqu'il s'agit de parties d'objets). Nous proposons donc ici un affinement de la relation *partie-de*.

4.1 De la relation de succession vs précédence à la relation de condition

Pour décrire un processus découpé en phases (prenons deux phases, X et Y), on peut avoir besoin de savoir au moins si :
- X précède Y
- X suit Y
- X se déroule en même temps qu'Y (concomitance)
- X a lieu pendant Y.

Ce sont ces informations que l'on va chercher à retrouver dans le corpus, par l'application de patrons de fouille définis avec Sato.
Dans la présentation qui suit, nous nous intéressons plus particulièrement à la relation de précédence vs succession.
On peut définir un patron très général faisant intervenir des marqueurs verbes comme *précéder, suivre, succéder* ou des marqueurs prépositions comme *avant* ou *après*.
L'application du patron verbal ne nous donne que de très faibles résultats, seuls deux contextes mettent effectivement en évidence la succession de deux activités.
Par contre, les prépositions font apparaître des données plus pertinentes. On a, par exemple : *La phase d'Architecture débute après approbation des Spécifications Fonctionnelles et Techniques Produit* ou *Le Dossier de Conception Produit résultant doit être accepté avant la phase d'Intégration Produit*. L'examen de ces deux exemples nous amènent à prendre en compte l'idée de condition que l'on peut considérer comme un affinement de la notion de précédence (une condition précède toujours une action). En effet, c'est *l'approbation des Spécifications* qui conditionne le démarrage de la *phase d'Architecture*, de même que l'*acceptation du Dossier de Conception* conditionne le début de la *phase d'Intégration.*
De plus, si on examine de plus près cette idée d'un déclenchement lié à une *approbation* ou *acceptation*, on met en évidence une règle générale propre au processus que l'on décrit dans notre corpus. En effet, si l'on recherche le marqueur de condition le plus intuitif dans le corpus, *condition$* (avec une troncature signifiant une variable en fin de mot), on trouve trois occurrences dont l'une décrit explicitement *le cycle de développement* comme *un enchaînement de phases* dans lequel *le passage d'une phase à la suivante est conditionné par l'acceptation de la phase précédente*.
Ainsi, à partir de marqueurs de succession/précédence, on a obtenu non pas des contextes où s'exprime cette relation mais des contextes où c'est la relation de condition qui est exprimée.
Si l'on décide de s'intéresser désormais à cette relation de condition, il convient de trouver des marqueurs suffisamment précis pour qu'ils génèrent le moins possible du bruit ou du silence.

4.2 La relation de condition

Ce paragraphe va nous permettre d'étudier plus en détail la relation de condition, en particulier, les marqueurs qui permettent de la mettre en évidence le plus sûrement ainsi que la meilleure façon de la modéliser.

4.2.1 Recherches des marqueurs de condition

Repartant d'une perception intuitive, on peut chercher si les marqueurs généraux de la condition se retrouvent dans le corpus. De ce point de vue-là, des verbes comme *conditionner, entraîner, provoquer* sont très peu productifs (à part l'exemple de *conditionner* que nous avons donné ci-dessus). Des marqueurs plus spécifiques du corpus comme *approbation* ou son équivalent *acceptation* sont eux trop généraux pour ne renvoyer que les contextes où une relation de condition est exprimée. En réalité, ces mots sont plutôt utilisés pour introduire des critères de Qualité (on trouve ainsi souvent : *procédure d'acceptation, modalités d'acceptation, critères d'acceptation, test d'acceptation)*. Par contre, ces deux mots étant des déverbaux, nous pouvons tester si la forme verbale n'apparaît pas dans les textes. Il s'avère, en effet, qu'une forme verbale (la forme participe passé) est très utilisée dans le corpus dans des contextes où une condition est exprimée. On trouve ainsi :
Phase de réalisation du produit est achevée lorsque tous les composants sont acceptés
Le bilan produit est effectué lorsque les produits sont acceptés.
Or, l'examen des contextes dans lesquels le participe passé *accept(é,ée,ées,és)* est utilisé (*approuvé* n'est pas utilisé) fait apparaître une corrélation très nette entre le fait qu'une condition est exprimée et le fait que le contexte d'apparition est une subordonnée temporelle. Il faut cependant une principale particulière puisqu'elle doit exprimer le début ou l'achèvement d'une phase.
Un pas supplémentaire nous montre d'ailleurs que la notion d'acceptation qui induit la notion de condition dans la subordonnée peut s'exprimer simplement par un participe passé même s'il est utilisé avec un autre verbe qu'accepter. Ainsi dans, *la phase de validation est achevée*

[5]L'étude de ce corpus est réalisée dans le cadre d'un projet financé par le GIS Sciences de la Cognition, " Terminologie, modélisation des connaissances et systèmes hypertextuels de consultation de documentation technique ".

lorsque les tests de validation ont été effectués avec succès, c'est bien l'aspect accompli apporté par le participe passé (ici au passif) *effectués* qui permet de comprendre qu'une condition a été respectée.
On pourrait dès lors en déduire que la configuration de marqueurs à utiliser pour repérer l'expression d'une condition dans ce corpus particulier serait : Conj. de subordination temporelle + verbe au participe passé.
Or, l'application de ce patron ramène beaucoup trop de bruit, d'une part parce que Sato ne contient pas d'analyseur syntaxique et que beaucoup de formes peuvent être considérées comme des participes passé, d'autre part parce que n'importe quel verbe au participe passé ne peut pas être considéré comme manifestant l'accomplissement d'une condition. Dans *lorsque les besoins ne sont pas couverts par les chapitres du document de référence, des chapitres supplémentaires peuvent être rédigés,* c'est peut-être une cause qui est indiquée par la subordonnée mais certainement pas une condition.
Le patron de fouille qui s'avère le plus efficace fait intervenir une notion de phase dans la principale et un marqueur de subordination temporelle (lorsque, dès, après avant, quand), sans contrainte a priori sur la subordonnée temporelle.
La notion de phase peut s'exprimer soit au moyen d'un déverbal (et nous recueillons ici les résultats de l'analyse des déverbaux) soit au moyen d'un équivalent du mot *phase* (équivalents que nous avons recherchés dans Sato par un examen des contextes contenant un démonstratif suivi d'un nom autre que *phase* ou une nominalisation, dans des contextes où ils constituent une reprise anaphorique de *phase*, comme *La phase de conception préliminaire (...). Cette tâche (...)*). Ainsi, en plus du terme *phase*, on a également retenu les termes *tâche, activité, processus*. Le patron de fouille qu'on obtient alors est le suivant :
*(tâche$,activité$,processus,revue$,phase$,$*Dev=dev)*.5 (lorsque,dès)*.*[6]
Avec ce nouveau patron de fouille, obtenu progressivement par adaptation en fonction des résultats renvoyés par Sato, les occurrences obtenues sont toutes très pertinentes, on en citera quelques exemples : *Les activités liées à l'intégration débutent dès que (...), Cette phase est achevée lorsque (...)*. Les occurrences nous donnent toutes, pour une phase déterminée dans la principale, la condition de son début ou de sa fin.

4.2.2 Modélisation de la relation de condition

A coté d'une relation de précédence/succession (et en lien avec elle) il est nécessaire de prendre en compte une relation de condition.
Un examen plus fin des résultats obtenus par application du patron de fouille montre que les conditions s'appliquent pour le début ou la fin des phases. Un schéma général serait donc :

X conditionne le début de Y
Z conditionne la fin de Y.

Si nous visons à construire une BCT, il est contre-intuitif de retenir comme terme des syntagmes comme *début de la phase Y* ou *fin de la phase Y*. Cette façon de procéder, si elle permettait de conserver une relation conditionne serait très peu économique puisque pour chaque étape, on aurait au moins deux termes concernés, celui qui indiquerait le début et celui qui indiquerait la fin.
Nous avons donc fait un choix plus économique qui consiste à retenir deux relations : *conditionne le début* et *conditionne la fin*.
Le schéma suivant montre un exemple de mise en oeuvre de ces deux relations.

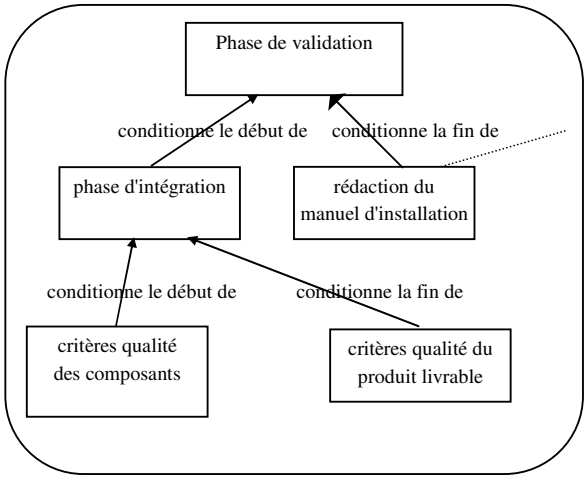

Figure 2 : Partie du réseau conceptuel dans lequel des concepts sont reliés par les relations « condition »

A partir de marqueurs supposés d'une relation, nous avons évolué vers l'identification d'une autre relation pour laquelle nous avons progressé, pas à pas vers l'affinement d'un patron de fouille très efficace.
Aussi bien la modélisation de cette relation que la définition de ce patron de fouille sont très spécifiquement dépendants du corpus à l'étude. Il n'aurait pas été possible d'imaginer de tel résultats avant l'expérimentation. Il n'aurait pas été possible non plus de les trouver sans une réflexion méthodique intégrant les résultats fournis par le logiciel à chaque étape.

Par une intégration méthodique des résultats fournis par Sato dans la définition de nouveaux patrons de fouille, nous progressons vers la définition d'un réseau conceptuel qui est soumis, dans une dernière étape, à un expert du domaine qui le valide. Alors seulement, la liste des termes et des relations retenus est arrêtée.

5. Conclusion

Si la constitution de bases de connaissances terminologiques ne peut ignorer les nombreux outils, d'extraction de termes candidats, d'analyse de corpus, etc. qui visent à l'assister, il reste cependant à préciser l'utilisation qu'on peut faire des résultats qu'ils fournissent. La méthode que nous avons décrite ici a pour objectif de proposer une façon d'intégrer ces outils dans une démarche linguistique de constitution de terminologies. Il apparaît clairement dans le processus que nous avons décrit que les outils interviennent essentiellement pour assister le linguiste-terminologue dans sa tâche, pour le guider dans son recueil et de ce fait qu'il est nécessaire que ces outils soient d'une grande

[6] Il faut lire ce patron de fouille de la manière suivante : il s'agit de rechercher tous les contextes dans lesquels un mot de la liste (phase, etc.) est suivi, dans un contexte de 0 à 5 mots près, d'une conjontion de subordination temporelle.

souplesse afin de laisser toujours possible des retours en arrière, au fil de la démarche.

6. Bibliographie

Bourigault, D. & Lépine P. (1994). Méthodologie d'utilisation de Lexter pour l'acquisition des connaissances à partir de textes. In Actes des Journées Acquisition Validation Apprentissage, JAVA-94, 1--13.

Bourigault, D. & Condamines, A. (1995). Réflexions sur le concept de Base de Connaissances terminologiques. In Actes des 5èmes Journées du PRC IA, Nancy 1-3 février 1995, Toulouse : Teknea, 425--444.

Borillo, A. (199). Exploration automatisée de textes de spécialité : repérage et identifcation automatique de la relation lexicale d'hyperonymie. LINX, 34-35.

Brunet, E. (1995). Un hypertexte statistique pour grands corpus : HYPERBASE. In Lexicomatique et Dictionnairiques, IVes Journées scientifiques du réseau Lexicologie, Terminonologie, Traduction, Lyon.

Condamines, A. & Amsili, P. (1993). Terminology between Language and Knowledge : An example of Terminological Knowledge Base. In TKE-93 : Terminology and Knowledge Engineering, Frankfurt: Indeks Verlags, 316--323.

Condamines, A. (1996a). Aide à l'acquisition des connaissances par l'étude de la terminologie. In N. Aussenac-Gilles, P. Laublet & C. Reynaud (Eds.), Acquisition et Ingénierie des connaissances : tendances actuelles (pp. 247--265). Toulouse : Cepadues.

Condamines, A. (1996b). Analyse de textes pour l'acquisition de données terminologiques. Terminologies Nouvelles, 14, 35--42.

Daoust, F. (1992). SATO (Système d'Analyse de Textes par Ordinateur) version 3.6, Manuel de référence. Centre ATO Université du Québec à Montréal.

Descamps, J.L., Mochet, M.A., Lewin, T., Lamizet, B. & Costes, D. (1992). Sémantique et concordances. Publication de l'INALF, Collection "St Cloud", Paris : Klincksieck.

Hearst, M.A. (1992). Automatic acquisition of hyponymys from large text corpora. In Actes du 15ème Colloque international en linguistique informatique, COLING-92, Nantes, 539--545.

Herviou, M.L., Ogonowski, A. & Dauphin, E. (1994). Tools for extracting and structuring Knowledge from Texts. In Actes du Colloque international en linguistique informatique, COLING-94, Japon : Kyoto.

Jackiewicz, A. (1996). L'expression lexicale de la relation d'ingrédience (partie-tout). Faits de Langues, 7, 53--62.

Jouis, C. (1995). SEEK, un logiciel d'acquisition des connaissances utilisant un savoir linguistique sans employer de connaissances sur le monde externe. In Actes des Journées d'Acquisition de Connaissances du PRC-GDR-IA du CNRS, Grenoble : 5-7 avril, 159--172.

Kavanagh, J. (1996). The Text Analyzer: A tool for extracting knowledge from text. Master thesis, Université d'Ottawa.

Meyer, I. & Mackintosh, K. (1996). The corpus from a terminographers viewpoint. International Journal of corpus Linguistics, 1 (2).

Otman G. (1996). Le traitement automatique de la relation partie-tout en terminologie. Faits de langue, 7, 43--52.

Rastier, F. (1995). Le terme : entre ontologie et linguistique. La Banque des mots, 7, 35--65.

Rebeyrolle, J. (1996). Définition et textes spécialisés : vers une modélisation. In Actes des Troisièmes Rencontres de l'Atelier des Doctorants de l'Université de Paris7, 9--11.

Rousselot, F., Frath, P. & Oueslati, R. (1996). Extracting Concepts and relations From Corpora. In Actes de ECAI'96, 12th European Conference on Artificial Intelligence, John Willy & Sons.

Séguéla, P. & Aussenac, N. (1997). Un modèle de base de connaissance terminologique. In Actes des 2Ëmes rencontres de Terminologie et Intelligence Artificielle : TIA'97.